Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing

Dirk Schübeler[1], David Scalzo[1], Charles Kooperberg[2], Bas van Steensel[3], Jeffrey Delrow[4] & Mark Groudine[1,5]

Published online 30 September 2002; doi:10.1038/ng1005

Replication of the genome before mitotic cell division is a highly regulated process that ensures the fidelity of DNA duplication. DNA replication initiates at specific locations, termed origins of replication, and progresses in a defined temporal order during the S phase of the cell cycle. The relationship between replication timing and gene expression has been the subject of some speculation[1]. A recent genome-wide analysis in *Saccharomyces cerevisiae* showed no association between replication timing and gene expression[2]. In higher eukaryotes, the limited number of genomic loci analyzed has not permitted a firm conclusion regarding this association. To explore the relationship between DNA replication and gene expression in higher eukaryotes, we developed a strategy to measure the timing of DNA replication for thousands of genes in a single DNA array hybridization experiment. Using this approach, we generated a genome-wide map of replication timing for *Drosophila melanogaster*. Moreover, by surveying over 40% of all *D. melanogaster* genes, we found a strong correlation between DNA replication early in S phase and transcriptional activity. As this correlation does not exist in *S. cerevisiae*, this interplay between DNA replication and transcription may be a unique characteristic of higher eukaryotes.

To investigate the relationship between gene expression and timing of DNA replication in higher eukaryotes, we developed a microarray-based method of constructing a genome-wide replication timing map. For this analysis, we used the *D. melanogaster* Kc cell line, which has been used in several microscopic studies of replication timing[3,4]. To label and isolate newly replicated genomic DNA, we adapted an established protocol to isolate cells in a defined part of S phase[5]. We labeled a non-synchronized exponentially growing cell culture with the nucleotide analogue bromodeoxyuridine triphosphate (BrdU), then sorted labeled cells on the basis of total DNA content by fluorescence-activated cell sorting (FACS). DNA from each sorted fraction was immunoprecipitated with an antibody specific for BrdU to enrich for those genomic sequences that were replicated during the labeling period (Fig. 1*a*,*b*). The immunoprecipitated DNA from fractions of the first third ('early') and last third ('late') of S phase was amplified by PCR and each fraction was labeled with a specific fluorescent dye. We hybridized both products simultaneously to a microarray containing 6,500 sequences, 5,543 of which represented previously described expressed sequences from *D. melanogaster*[6]. A stronger fluorescence signal for one of the two dyes indicated enrichment of a sequence in one of the fractions (Fig. 1*c*). To ensure reproducibility, we carried out three independent experiments, each using a different passage of cells.

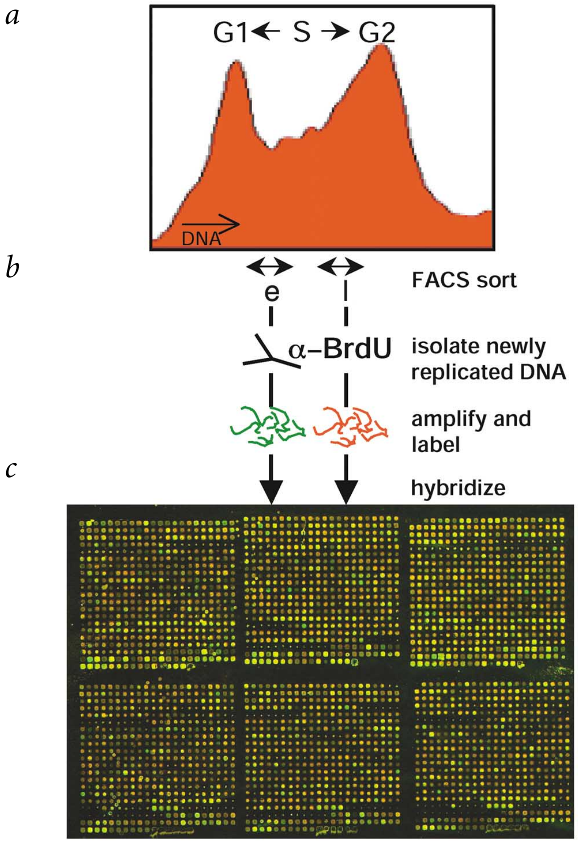

Fig. 1 Experimental strategy to measure the timing of DNA replication using spotted microarrays. ***a***, Cell-cycle profile of *D. melanogaster* cells after pulse-labeling with BrdU and staining of DNA with propidium iodide. Cells between the G1 and G2 peaks are in S phase. Two fractions, reflecting the early (e) and late (l) stages of S phase, were sorted by FACS. ***b***, After purification from the sorted cells, we enriched DNA containing BrdU by immunoprecipitation with an antibody specific for BrdU, and then carried out PCR amplification and fluorescent labeling. ***c***, Early and late fractions were hybridized simultaneously to a spotted microarray. A section of one of the array hybridizations is shown. Enrichment of a sequence in early or late fractions is indicated by green or red fluorescence, respectively.

[1]Division of Basic Sciences and [2]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [3]Netherlands Cancer Institute, Amsterdam, the Netherlands. [4]DNA array facility, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [5]Department of Radiation Oncology, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to D.S. (e-mail: dschubel@fhcrc.org) or M.G. (e-mail: markg@fhcrc.org).

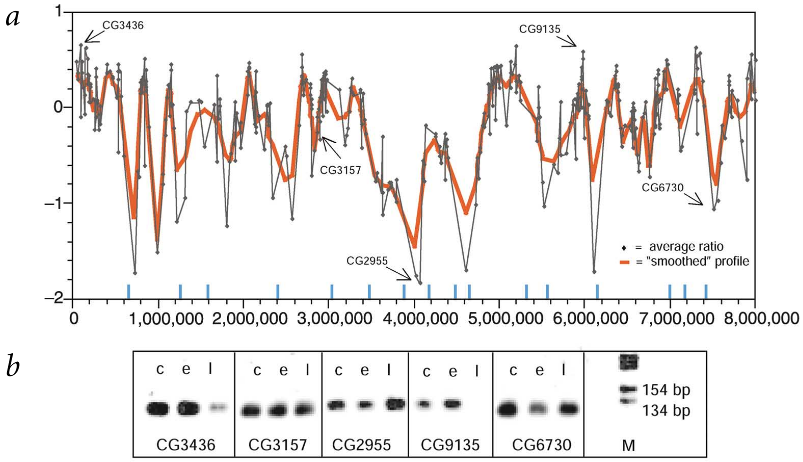

Fig. 2 Replication timing of genes in the part of chromosome 2L proximal to the telomere. ***a***, Raw and processed data for 386 probes, which span over 8 Mb in this particular region. Shown is the average ratio from six experiments for each individual spot, and the resulting replication profile after local polynomial smoothing (loess). Genomic distances greater than 100 kb between neighboring spots are indicated with bars above the *x* axis. *y* axis, ratio of array signals (log$_2$ (early/late)); *x* axis, chromosomal position (bp). ***b***, Replication timing of specific genes as measured by PCR. We amplified sequences from control genomic DNA (c) and the early (e) or late (l) fractions before hybridization. The products from five sequences in the region displayed in Fig. 2*a* are shown. In each case, PCR confirmed the array observations for a particular sequence. Genes CG3436 and CG9315 showed early replication and were enriched in the early fraction; CG2955 and CG6730 showed late replication and were enriched in the late fraction; and CG3157 showed an intermediate replication and was almost equally abundant in the early and late fractions. The gene names correspond to their entries in FlyBase. M is a size marker.

After each hybridization, the fluorescence signal obtained from the early fraction was divided by that from the late fraction. For 5,221 sequences spotted onto the microarray, we obtained a measurement from the timing assay and we obtained a unique genomic position through a BLAST search against the *D. melanogaster* genome. The average chromosomal distance between sequences in this set was 20.5 kb. These 5,221 sequences provided a genomic scaffold for the generation of a chromosome-wide profile of replication timing. Timing ratios were log$_2$ transformed, so that more positive values indicated earlier replication and more negative values indicated later replication.

We used semi-quantitative PCR controls for a subset of genes to confirm the enrichments seen in the array hybridizations (Fig. 2*b*). The timing ratios determined from the array hybridizations were aligned on the chromosomal arms according to the genomic position of the sequences (Fig. 2*a*). Neighboring sequences tended to have comparable ratios, indicating that they were replicated at similar times in S phase. We vali-

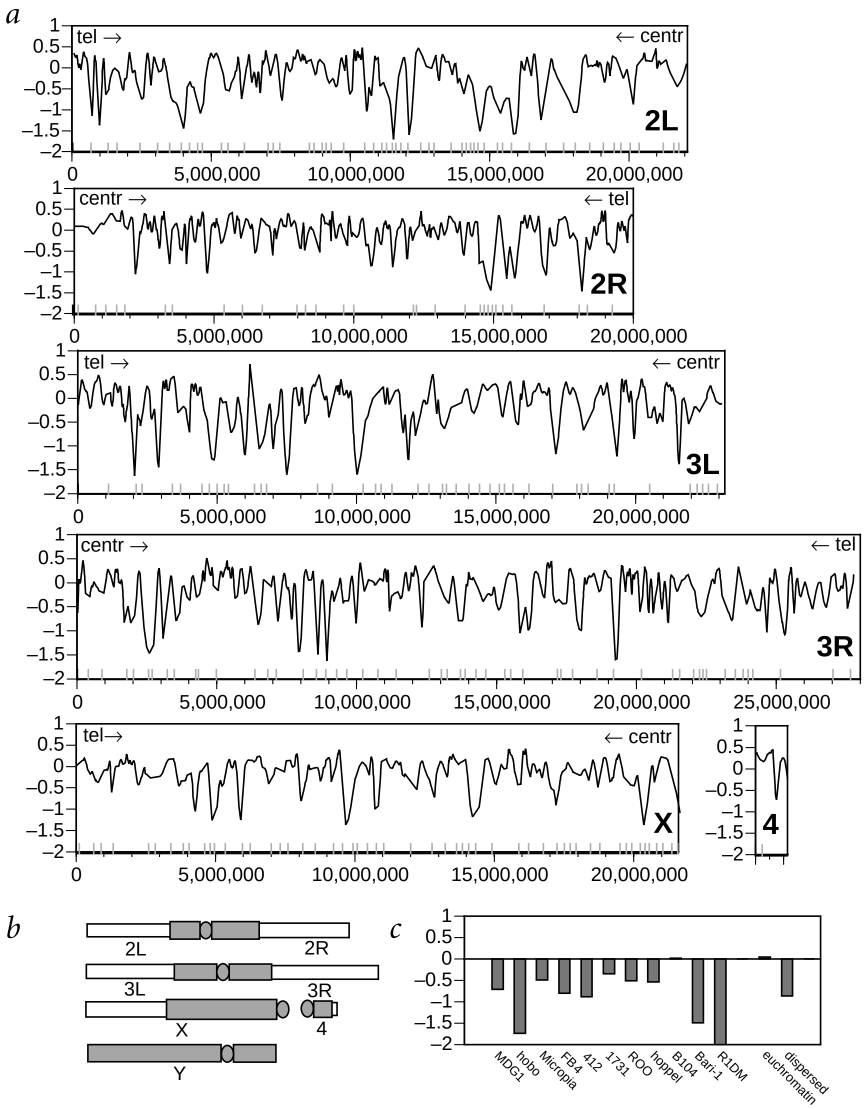

Fig. 3 Replication profile for the sequenced *D. melanogaster* genome. ***a***, Profile of the processed data for the euchromatic arms of the *D. melanogaster* chromosomes. Bars above the *x* axis indicate distances greater than 100 kb between neighboring spots. ***b***, Mitotic chromosomes of *D. melanogaster*. The sequenced euchromatin is shown in white and the heterochromatin in gray. ***c***, Replication timing ratio for dispersed transposable elements, which showed significantly later replication timing ($P < 10^{-6}$; 11 randomly chosen sequences of the unique 5,221 euchromatin spots had a higher average replication timing ratio than the 11 transposable elements shown here in all 1,000,000 samples). Of the transposable elements shown, Bari-1 and R1DM are localized exclusively to heterochromatin. They replicated later than 98% of the euchromatin sequences on the array, indicating late replication of heterochromatin. The other transposable elements listed are present in euchromatin and heterochromatin, and thus their replication timing represents the average of euchromatin and heterochromatin integration sites. Euchromatin, average of single-copy genes; dispersed, average of dispersed elements shown.

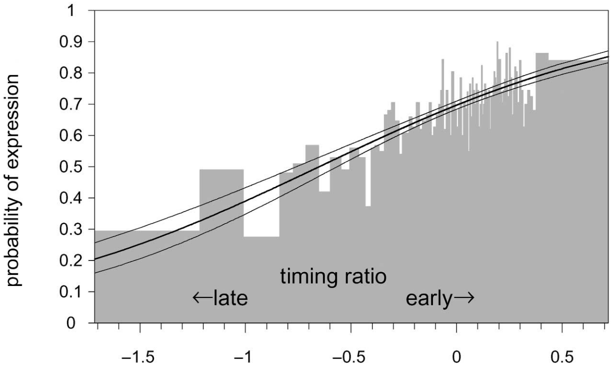

Fig. 4 Relationship between transcription and timing of DNA replication. To determine whether timing of DNA replication and expression status of a gene were correlated, we analyzed the expression status of all genes present on the array by cDNA hybridization. Genes were scored as either expressed or not expressed. Of 5,077 sequences, 3,466 were expressed and 1,611 were not expressed. Genes with a unique genomic position were ranked for plotting purposes according to their replication timing as measured in the array experiments and divided into groups of fifty (based on similar replication timing). Subsequently, the percentage of expressed genes (that is, the probability of expression) was calculated for each of these groups. The percentage of expressed genes (*y* axis) is plotted versus the replication timing (*x* axis), with each column representing one group. This representation shows that the percentage of expressed genes was higher for genes that are replicated early than for genes that are replicated late. The relationship between the probability of expression and replication timing was further analyzed using logistic regression, which is a statistical method used if there are only two potential outcomes for one of the two variables (in this case, a gene is either expressed or not expressed). The resulting logistic regression curve (thick line) and 95% confidence intervals (outer lines) showed a strong correlation between early replication and transcriptional activity as well as between late replication and lack of transcription (slope of 1.28 with a *t* statistic of 14.1 and a *P* value of10^{-44}).

dated this visual observation using an autocorrelation analysis to test whether the timing ratios of neighboring sequences were related. For all chromosomal arms, we observed a correlation over an average distance of at least five and up to eight neighboring sequences (see Web Figure A online). Thus we concluded that, on average, neighboring sequences were replicated at similar times during S phase. We used local polynomial smoothing ('loess')[7] to create a complete profile from the raw data set. This mathematical model reduces underlying noise and outliers that may result from cross-hybridization to sequences spotted on the array or from inaccuracies in the genome annotation. Comparable data processing was used in the recent analysis of the timing of genome replication in yeast[2]. This procedure resulted in a fitted curve by removing noise in regions with high probe density (Fig. 2*a*). Similar processing was used for the remainder of the sequenced *D. melanogaster* genome to generate a genome-wide profile (Fig. 3*a*). It should be noted, however, that the precision of the map is influenced by the genomic density of spotted sequences, and the use of DNA arrays with higher probe density may generate a more detailed map of replication timing for those chromosomal regions with low probe coverage.

The sequenced portion of the *D. melanogaster* genome represents the cytogenetically defined euchromatin, where the vast majority of genes reside. Euchromatin can be distinguished from heterochromatin, which comprises primarily pericentromeric regions and the Y chromosome (Fig. 3*b*). Although studies in various systems have suggested that euchromatin located proximal to heterochromatin replicates late in S phase[8], our analysis of the complete *D. melanogaster* euchromatin showed no such trend for sequences proximal to either telomere or centromere (Fig. 3*a*).

Previous microscopic studies of DNA replication in *D. melanogaster* Kc cells have established that *D. melanogaster* heterochromatin replicates later than euchromatin[3,4,9]. The lack of sequence information and high repetitiveness make heterochromatic sequences refractory to DNA array analysis. The array used in our study contains a number of interspersed transposable DNA elements, however, some of which are localized exclusively to heterochromatic regions. Most interspersed elements showed a negative ratio for their replication timing, and their average replication timing was significantly later than that of the unique euchromatic sequences (Fig. 3*c*). More importantly, the two classes of transposons (Bari-1 and R1DM) that reside exclusively in heterochromatin in *D. melanogaster*[10,11] replicated later than 98% of the euchromatin. These results support previous observations that pericentromeric heterochromatin replicates late in S phase, and are compatible with the idea that the establishment and maintenance of a repressive chromatin structure may be linked to late replication in S phase. Notably, the region of chromosome 2L proximal to the centromere did not replicate late in S phase, even though it contains genes, such as light (*lt*) and concertina (*cta*), that require proximity to heterochromatin for proper regulation[12] and therefore have been defined as heterochromatic genes. These genes, however, reside in β-heterochromatin, which, unlike the centromeric α-heterochromatin, does not possess highly repetitive DNA and has a euchromatin-like gene density[13]. Thus, the various forms of *D. melanogaster* heterochromatin differ in several respects, including replication timing. Regions of late replication were also interspersed throughout the cytogenetically defined euchromatin (Fig. 3*a*), and further studies will be required to determine whether such sequences reside in a repressed and/or potentially heterochromatic structure.

It has been widely speculated that replication early in S phase allows the packaging of newly synthesized DNA into an 'open' chromatin structure and thereby mediates transcriptional activity. This hypothesis predicts a correlation between expression and replication timing. Such a correlation has been observed for a few genes in higher eukaryotes[14,15] but excluded for others[16,17].

To analyze the relationship of replication timing and gene activity on a genomic scale, we determined the transcriptional activity of all genes present on the array in the same cell line and compared expression status with replication timing for each individual gene. We obtained replication timing and expression data for 5,077 genes on the array. Of these, 3,466 were expressed and 1,611 were not expressed. A comparison of expression and replication timing for this set, representing over 40% of all *D. melanogaster* genes, showed a correlation between expression and the time of DNA replication (Fig. 4). The percentage of replicated genes that were expressed gradually decreased from early to late S phase. Consequently, early replication in S phase coincided with a higher likelihood of gene activity on a genome-wide level, suggesting a concerted regulation of transcription and the timing of DNA replication for a large subset of genes in *D. melanogaster*. This result contrasts with a recent study in budding yeast in which no correlation between transcription and timing of DNA replication was observed[2].

Taken together, these results suggest that the interplay between transcription and DNA replication might be a unique feature of higher eukaryotes. Such co-regulation may, for example, reflect a more complex nuclear compartmentalization[8,18] or a regulation of replication initiation use[19] associated with tissue-specific gene regulation. If so, further studies using genome-wide profiling of DNA replication in differentiated cells may provide important insights into the concerted regulation of gene expression and replication timing.

Methods

Tissue culture and BrdU labeling. We cultured *D. melanogaster* Kc cells as described[20]. For labeling of newly synthesized DNA, we added BrdU (Sigma) to a logarithmically growing culture at a final concentration of 50 μM. After 60 min of incorporation time, cells were washed two times in cold PBS, resuspended in 2.5 ml PBS, fixed by slowly adding 7.5 ml of cold ethanol and stored at –20 °C.

FACS sorting and DNA immunoprecipitation. We prepared cells and carried out FACS as described[15] with the following minor modifications. We used propidium iodide to stain cellular DNA and sorted cells into S-phase fractions on the basis of DNA content (Fig. 1). We sorted using two gates representing roughly the first and last thirds of S phase. We collected 40,000 cells from each fraction directly into lysis buffer without salmon sperm DNA. DNA was purified as described[15], sonicated, denatured and immunoprecipitated with a monoclonal antibody specific for BrdU (Becton Dickinson).

PCR amplification and fluorescent labeling of BrdU enriched DNA. We amplified the denatured and immunoprecipitated DNA as described[21] with minor modifications. Briefly, we first created templates for PCR by randomly annealing and extending a 3′ degenerate primer using T7 polymerase. We then used primer B, representing the T7-extended, non-degenerate portion of primer A, to amplify the templates. Primer B was labeled with either Cy3 or Cy5 at the 5′ end (Qiagen Operon). Size distribution and fluorescence of the PCR product was confirmed by agarose gel electrophoresis followed by a fluorescence analysis for Cy3 and Cy5 using a fluorescence scanner (Molecular Dynamics). We carried out four PCR reactions for each of the cell fractions (early and late), which were processed as described[20] and pooled before the hybridization. We did array hybridization and washing as described[20]. We carried out three independent experiments of BrdU incorporation and subsequent array hybridizations. In addition, to account for the potential influence of the fluorescent dyes on the hybridization results, we amplified and hybridized immunoprecipitated DNA from each experiment twice; the only difference between hybridizations was the switching of the fluorescent dye labels.

Control PCR. We designed primers to amplify products of 80–120 bp to control for abundance in the early or late fraction by PCR. We carried out PCR using roughly 10 ng of DNA, standard PCR conditions and 27 cycles of amplification. PCR products were separated by agarose gel electrophoresis and visualized by ethidium bromide staining of the gel. Primer sequences are available upon request.

Array analysis. We analyzed fluorescence scans of the arrays using the Genepix software package (Axon Instruments) with improved background correction for the replication timing arrays[22]. The ratio of the two fluorescent dyes from each hybridization was $\log_2$ transformed and normalized using intensity dependent normalization[23] using loess[7] to correct for nonlinearities. We used the average value of the six hybridizations for further analysis of the replication timing data set. The median s.d. for this set of sequences was 13.7%.

Expression profiling. After isolating total RNA from a logarithmically growing culture of cells using Trizol following the manufacturer's (Gibco) recommendation, we used 30 μg RNA for cDNA synthesis and labeling as described[24]. The labeled cDNA was hybridized to the array and processed similarly to the genomic hybridizations. A gene was scored as positive for expression if the fluorescent intensity of the spot was greater than 2.5 times the average background in both channels in at least two of three independent experiments.

Microarray preparation. Owing to the lack of genomic microarrays for *D. melanogaster*, we used an array assembled by the Northwest Flychip consortium instead. This array represents a previously described set of sequences[20], but contains an additional collection of 5,543 expressed sequence tags from *D. melanogaster*[6].

Genomic coordinates of array probes. We compared the sequences of all expressed sequence tags present on the microarray against the *D. melanogaster* genome (BDGP Release 2) using MEGABLAST (National Center for Biotechnology Information). For 5,475 expressed sequence tags, we found a single significant match. We approximated the chromosomal positions by using the coordinate halfway between the most 5′ and 3′ matching genomic nucleotides.

Data processing. We used the local polynomial smoother loess[7,25] to generate a smoothed replication profile. Local polynomial smoothing is a generalization of local averaging methods (kernel), but has a higher order of accuracy. The bandwidth was estimated separately for each chromosome using 'leave-one-out' cross-validation: that is, the smoothed replication profile was generated using all spots except one. This was done separately for each spot using various bandwidths, and we used the bandwidth that created the smallest mean squared prediction error for all spots.

URL. The complete data set can be accessed at http://parma.fhcrc.org/DSchubeler.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

We thank A. Oryan for Kc cells; E. Giniger for coordinating the assembly of the Northwest Fly Consortium microarray; C. Neal for array construction and cDNA analysis; J. Schreiber, D. Cimbora and members of the Groudine lab for suggestions; and S. Henikoff, D. Gottschling and M. Lorincz for comments on the manuscript. This work was supported in part by US National Institute of Health grants to M.G. and C.K., a grant from the Netherlands Academy of Sciences to B.v.S. and a fellowship grant from the Rett Syndrome Research Foundation to D.S.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 22 April; accepted 21 August 2002.

1. Cimbora, D.M. & Groudine, M. The control of mammalian DNA replication: a brief history of space and timing. *Cell* **104**, 643–646 (2001).
2. Raghuraman, M.K. *et al*. Replication dynamics of the yeast genome. *Science* **294**, 115–121 (2001).
3. Ahmad, K. & Henikoff, S. Centromeres are specialized replication domains in heterochromatin. *J. Cell. Biol.* **153**, 101–110 (2001).
4. Dolfini, S., Courgeon, A.M. & Tiepolo, L. The cell cycle of an established line of *Drosophila melanogaster* cells *in vitro*. *Experientia* **26**, 1020–1021 (1970).
5. Hansen, R.S., Canfield, T.K., Lamb, M.M., Gartler, S.M. & Laird, C.D. Association of fragile X syndrome with delayed replication of the *FMR1* gene. *Cell* **73**, 1403–1409 (1993).
6. Rubin, G.M. *et al.* A *Drosophila* complementary DNA resource. *Science* **287**, 2222–2224 (2000).
7. Cleveland, W.S. & Devlin, S.J. Locally-weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83**, 596–610 (1988).
8. Gilbert, D.M. Replication timing and transcriptional control: beyond cause and effect. *Curr. Opin. Cell Biol*. **14**, 377–383 (2002).
9. Sullivan, B. & Karpen, G. Centromere identity in *Drosophila* is not determined *in vivo* by replication timing. *J. Cell. Biol.* **154**, 683–690 (2001).
10. Caizzi, R., Caggese, C. & Pimpinelli, S. Bari-1, a new transposon-like family in *Drosophila melanogaster* with a unique heterochromatic organization. *Genetics* **133**, 335–345 (1993).
11. Jakubczak, J.L., Zenni, M.K., Woodruff, R.C. & Eickbush, T.H. Turnover of R1 (type I) and R2 (type II) retrotransposable elements in the ribosomal DNA of *Drosophila melanogaster*. *Genetics* **131**, 129–142 (1992).
12. Wakimoto, B.T. & Hearn, M.G. The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome 2L of *Drosophila melanogaster*. *Genetics* **125**, 141–154 (1990).
13. Eberl, D.F., Duyf, B.J. & Hilliker, A.J. The role of heterochromatin in the expression of a heterochromatic gene, the rolled locus of *Drosophila melanogaster*. *Genetics* **134**, 277–292 (1993).
14. Hatton, K.S. *et al*. Replication program of active and inactive multigene families in mammalian cells. *Mol. Cell. Biol.* **8**, 2149–2158 (1988).
15. Cimbora, D.M. *et al*. Long-distance control of origin choice and replication timing in the human β-globin locus are independent of the locus control region. *Mol. Cell. Biol.* **20**, 5581–5591 (2000).
16. Hansen, R.S., Canfield, T.K. & Gartler, S.M. Reverse replication timing for the *XIST* gene in human fibroblasts. *Hum. Mol. Genet.* **4**, 813–820 (1995).
17. Smith, Z.E. & Higgs, D.R. The pattern of replication at a human telomeric region (16p13.3): its relationship to chromosome structure and gene expression. *Hum. Mol. Genet.* **8**, 1373–1386 (1999).
18. Kennedy, B.K., Barbie, D.A., Classon, M., Dyson, N. & Harlow, E. Nuclear organization of DNA replication in primary mammalian cells. *Genes. Dev.* **14**, 2855–2868 (2000).

19. DePamphilis, M.L. Replication origins in metazoan chromosomes: fact or fiction? *Bioessays* **21**, 5–16 (1999).
20. van Steensel, B., Delrow, J. & Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.* **27**, 304–308 (2001).
21. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein–DNA association. *Nat. Genet.* **28**, 327–334 (2001).
22. Kooperberg, C., Fazzio, T.G., Delrow, J. & Tsukiyama, T. Improved background correction for spotted DNA microarrays. *J. Comput. Biol.* **9**, 55–66 (2002).
23. Yang, Y.H., Dudoit, S., Luu, P. & Speed, T.P. Normalization for cDNA microarrays data. in *Microarrays: Optical Technologies and Informatics* (eds Bittner, M.L., Chen, Y., Dorsel, A.N. and Dougherty, E.R.) (International Society for Optical Engineering, San Jose, California, 2001).
24. Pritchard, C.C., Hsu, L., Delrow, J. & Nelson, P.S. Project normal: defining normal variance in mouse gene expression. *Proc. Natl Acad. Sci. USA* **98**, 13266–13271 (2001).
25. Fan, J. & Gijbels, I. *Local Polynomial Modelling and its Applications* (Chapman-Hall, London, 1996).